Elimination of Redundant Information for Web Data Mining

Shakirah Mohd Taib, Soon-Ja Yeom, Byeong-Ho Kang
School of Computing, University of Tasmania, Australia
{mohds, s.yeom, byeong.kang}@utas.edu.au

Abstract

These days, billions of Web pages are created with HTML or other markup languages. They only have a few uniform structures and contain various authoring styles compared to traditional text-based documents. However, users usually focus on a particular section of the page that presents the most relevant information to their interest. Therefore, Web documents classification needs to group and filter the pages based on their contents and relevant information. Many researches on Web mining report on mining Web structure and extracting information from web contents. However, they have focused on detecting tables that convey specific data, not the tables that are used as a mechanism for structuring the layout of Web pages. Case modeling of tables can be constructed based on structure abstraction. Furthermore, Ripple Down Rules (RDR) is used to implement knowledge organization and construction, because it supports a simple rule maintenance based on case and local validation.

1. Introduction

Nowadays, a great number of web documents are created and published to the Internet world with a variety of styles and information for users. The content from Web documents has lots of information from different types of fields and they are designed for diverse groups of aimed users. Thus, Web content's extraction and analysis become more challenging due to the growing number of online accessible Web documents and the density of the content's organization. The pages use a more complex structure for the design and present their information in various Web styles using numerously available Web editors. Web content's extraction and analysis are very important processes in the information retrieval system, Web classification and monitoring system. Due to the current trend of layout design of Web pages, most information is divided in according to contents' nature: navigation, main content, auxiliary information. Out of the various layout practices, the more accurate detection of main content area means the more accurate classification or monitoring of information.

For Web classification, many people use techniques that retrieve and extract the data from Web content and classify them based on keyword categorization. However, some researchers have done their work on the Web structure analysis scope for Web documents classification. Their works related to research on Web table structure and representation of HTML hierarchy. Based on the analysis, an abstraction of layout structure can be done and it makes the interpretation of Web layout structure easier. Abstraction knowledge can be represented in many ways. Common type of representation of abstraction knowledge is "rule-type" that makes the knowledge become generalized to different situations. This study contributes an analysis-based approach to Web's structure understanding for Web document's section classification. This work also produces a case model with a set of rules that represent abstraction knowledge of Web's structure. The case modeling is used for classification which is one of the important parts in section monitoring systems.

2. Literature Review

2.1 Web Classification

Web documents are more difficult to analyze than normal text documents, because they use evocative language instead of using descriptive languages [10]. Therefore, it contains some noises that make the content's detection more complex and requires sophisticated techniques to enhance the effectiveness of document classification.

Most of the users need to focus on the information that fits their interest only and some unsuitable content has to be ignored or filtered out. It is believed that 99% of the information on the Web is of no interest to the 99% of the people [2]. Unsuitable contents on Web pages that provide a unrelated information to users can be detected by monitoring systems that check predefined Web pages and prompt users when there are changes in these pages [6].

Many researchers implement MCRDR knowledge acquisition for classification, filtering as well as monitoring systems [6, 7]. As MCRDR is successfully

implemented for creating knowledge base for Web information management, the idea of case based knowledge is suitable for creating a model for Web documents to improve the classification process.

2.2 Detecting structure in HTML documents

There are several techniques that can be used to detect layout structure in HTML documents. Tables and frames are two components that are commonly used to organize the contents of Web documents. Thus, a table or frame detection technique is required to give a view of Web page layout for an analyzing process. Many developers prefer to use tables rather than frames to design the Web page's layout. Even though frames can provide added context and consistency during navigation, they have several serious problems that are related to screen real estate, page model, the speed of the display and the complexity of web design [8]. Therefore, several researchers have reported their work in table mining due to the efficiency and the popularity of tables for Web page layout structure [1, 3, 4, 11, 12, 13].

Based on the common structure of Web page layout, the most important section of a Web page can be detected because many Web pages have a similar structure in terms of screen width, and small font size to fit into only one page. They also create a wider column in the middle page for main text. These factors were mentioned by Palme and he defined three main sections on common Web pages; main text in the middle column, navigation information on the left and additional reading connected to main text in the right column [5].

2.3 Section monitoring and table mining

Users need to classify which section has the most important content to be viewed. Some sections in Web pages are only for irrelevant information to the main topic of user's need: advertisements, unimportant links, pictures and other information that are unrelated to the user's interest [10]. Every section has its own characteristics that need to be identified in order to monitor particular sections containing useful information. As the table is the main element of Web page layout design, table understanding in Web documents is the most important step in section identification process. Tables are constructed to give a structure for Web page layout and to present information in an efficient way. There are two types of tables that have been defined; genuine tables are used for conveying the logical relations among the cells and non-genuine tables only used as a technique for categorizing the Web contents for simple viewing only [11].

Many researchers have reported on their work on Web table detection. However, most of them were focused on extracting information from tables in Web documents, thus they did more testing on genuine tables instead of non-genuine tables [4, 12, 13]. For the section monitoring, both types of tables are considered to be examined because the purpose of identifying a specific section to be monitored is not only applied for the data from specific tables in Web documents but for the important part from the whole contents. Yang and Luk [12] developed a framework for comprehensively analyzing the structural aspects of Web table and gave an explanation of how rules can be derived to classify a given tables. This mining process starts with understanding of the HTML tables and followed by table recognition by filtering the HTML texts before analyze the structure of the tables.

There has been other research on Web table mining that proposed a systematic way to mine tables from large scale of HTML texts [1]. They discussed table filtering, table recognition, table interpretations and application of table extraction. Therefore, they employed the tags and information in table cells as main components for recognizing and interpreting the tables.

There are some issues involved if complexities of tables are to be utilized [4]. Internal cell structures in some cases are used to provide internal structure instead of delimitating a single cell in the table. Some other issues are split cell, errors that mostly happened when developers using HTML tool to position document elements, omissions in HTML, constraint that related to rendering algorithm and reconstruction HTML with incorrect tables.

3. Implementation

3.1 The analysis of Web document layout

The objective of an analysis of Web document layout is to specify the common structures of Web layout that can be found in various online Web documents. Using this analysis, it is possible to define a standard area in a page such as the header, footer and main section that contains the most important information out of the other sections in Web documents. The most important part for Web page layout analysis is the structure analysis. It has been used as a base for Web documents content extraction program and information mining. It is the best way to segment the Web pages into sections or zones that can be classified into certain groups. Since tables are used frequently in Web documents, the structure analysis focuses more on table understanding and its interpretation based on a table tree or its hierarchy.

Although many Web designers design pages according to similar standard principles, some of them prefer to present their Web page in a different style of layout.

Thus, it creates a problem of non-standard Web document's style and construction.

The table understanding task has been grouped into two sub tasks; table detection and table decomposition. Every page has a number of sections or zones of contents. A zone content's classification presents as a second important role in Web document layout analysis. It is depends on the successful table detection method that was used as the first step in the analysis. There are two common strategies for zone content's classification. First strategy uses the rule-based/grammar-driven algorithm and another one uses statistical-based approach. We choose a rule-based approach which uses a set of rules of the grammar to derive decisions of monitoring sections in Web pages. We use features in HTML documents to construct rules for classification. These features are derived from HTML feature's analysis including the table structure detection and interpretation.

3.2 Table detection

Tables are the most flexible layout system for designing Web documents. However, as the table gets larger, it becomes more complex and sophisticated, and the code becomes even more difficult to read. The structure of Web page layout is a construction of HTML elements that can be presented as a hierarchical model. This is because complex tables are usually created as nested tables. The table-tree detection is done by extracting the HTML features from each document.

In our study we need to find features in both groups of tables, genuine and non-genuine, that provide significant separation between the important section and the unimportant section that has to be monitored. One of the processes in table detection is detecting features of each table in the structure. We use HTML tags analysis to extract the attributes from HTML documents. There are various kinds of tags and variations that can be used to create a layout of Web pages. Researchers who have done their table mining and classification were concerned about the use of <TR> and <TD> tags to count the average number of rows and columns in a table using a complex table ground truth program. However, we do not compute that layout features as we can compute another simple attributes to detect significant zones or sections. We measure the cell length in terms of number of characters.

Another group of features is the content type feature. The content within the table can be many different types of content such as images, hyperlinks, forms, strings, scripts, etc. Normally, the main section of the layout for an informative Web page is more likely to contain alphabetical strings than images or forms and fewer hyperlinks. Thus, they can be differentiated from the other sections that provide other information which is not important for common users.

The third group extracted from HTML tags analysis is keyword features. We identify several groups of common keywords that can be found from common news Web pages layout. Keywords are still important although we focus on structure classification because keywords can be assigned as the values of some attributes such as class name and table id.

3.3 Section monitoring of layout classification

We collect a set of data within a specific domain and choose informative or news Web pages as a domain of data collection for this study. Then we extract the table rank and composition and generate a table tree that presents the tables in a hierarchy structure. Many of the Web pages have the same pattern of layout but they use a different number of tables as the structure and use a different arrangement of the tables' hierarchy. Therefore, we also search a group of keywords that can be identified in the important sections or some keywords that are not relevant to the users. We have discussed about detecting the other features in the previous section.

3.4 Structure abstraction

In a news-Web assessment, we can abstract all the Web document structure into a number of patterns. This abstraction can be used to manage with the complexity of the structures which are using various ways of table construction. The next task is a task of matching the abstracted case against the decision knowledge. The assessment criteria will be compared to the abstracted structured available. Then a decision is the result of the matching. Abstraction knowledge can be represented using a "rule type" [9]. The rule expressions are depending on the value of the operand which set to be the standards for numeric or non numeric attributes.

In a Web section monitoring domain, the decision rules are not simple. It cannot be the final decision by getting true for all criteria for a certain case. If the table construction is simple and used less than 10 tables, we can use a simple rule that matches easily and achieve a correct decision. However if more tables are used, then more rules have to be matched. Otherwise the wrong section of layout will be monitored. Therefore, we come out with another simple solution by comparing the cell length or character size for each table. The largest cell or table that has a maximum number of characters is defined as the main and important section among the others.

3.5 Case modeling approach for classification

We use the RDR approach to maintain the consistency of rule creation. It is a simple way of acquiring knowledge accessible to an end-user.

Experimental methods used in this study are divided into two main categories.

- Pre-processing
  - collect a set of cases (i.e. Web documents)
  - extract attributes from HTML data
  - create empty RDR knowledge base
- Processing
  - get the next case
  - get a conclusion drawn by rules
  - ask the expert, if agreed with the drawn conclusion, get another case and back to the first step.
  - if the expert doesn't agree with the conclusion, identify the last satisfied rule and attach a new rule to the true or false branch of the last evaluated rule.
  - back to the beginning

We extract a number of attributes during the pre-processing stage. However, some of the attributes were identified from new cases during the processing stage. The extracted attributes are from structure features such as links and heading tags and word group features including the size of characters and keywords groups. Keyword detection is also important in the experiment. In spite of extracting the structure attribute from the HTML source, we also search for a number of keywords that are essential to give an expression about attributes.

4. Results and Discussion

4.1 Analysis of common layout of Web documents

There are about five common layout structures that are usually used by news Website designers. The most popular layout has three columns or vertical sections and two horizontal sections for the header and footer. This layout has the main text in the middle column, navigation or menu information in the left column and additional information related to main column in the right column. However, some designers had tried to have different arrangement by placing the navigation menu in the right column or in some other ways. For this study, we collected 67 pages and 28 Web pages (42%) are categorized in this pattern. The minimum number of tables that were used to create the structure of this pattern is only three tables but it can be more than 90 tables or greater if the designers create many sub sections and tables for images. 33 pages out of 67 have tables of somewhere between 7 and 51 and are evenly distributed. From this analysis, we can say that it is difficult to classify Web pages into different patterns of layout based on the number of tables composed for layout structure.

4.2 Different classification factor

Another analysis that has been done in this research is classification by ranks of a table tree. Most of Web pages have nested tables in their structure. Those tables can be represented as a tree and we can categorize them by their ranks or levels. Many designers used only one or three tables in the first rank. One table is used as a main frame to the whole structure and the second level has three or four tables. They used three tables to assign the top table as a header or banner section, the second table as main text and the third table as a footer that has usually a copyright and contact information section. Normally, there are no more nested tables in the first and third tables, only the second table which is the main section contains some other tables for other sub sections. Most of Web pages used for this experiment have the maximum size of characters for their main text column or section. However, some sub sections in that column are not supposed to be monitored because of the highly possible cases of advertisement, poll or vote section or unimportant list of hyperlinks. Classification by the number of table in first rank and maximum length of character are two basic elements for rule's construction.

4.3 Rules and case model

Experiments have been done to 2,276 tables from 65 Web pages. These experiments used a simple program that detect tables from Web pages and classify them into a table tree structure. After that, it records the attributes from HTML documents in a database and uses them to check with the available rules created. We as the experts have to decide whether the table is correctly classified into their categories or the decision is correct. For this experiment, only two conclusions are defined; monitor or do not monitor the section or specifically conclude for every table in the Web page structure. The main purpose of choosing RDR is to ensure that any modifications leave the rule consistent.

There are several steps used in this experiment. The experiment was started from collecting data for training and testing and ending up with a set of accuracy graph and a set of rules presented as RDR tree structure. From the table tree, the program compares the ranks between the tables collected. First rank tables are not necessary to be trained because most of Web page's designers set three or four tables in first rank of the table's structure. Size of

characters in each table is one of the main attributes used for building a started rule in this experiment. Most of the attributes are collected during pre-processing and some are added after the rules are modified.

Keywords are the collection of some words that commonly used for section headings. They are represented in a strong or bigger font size than other isolated text in the Web page. They can be detected in the summary columns or tables within middle section in most of news Web pages. Another similar attribute is an image link which is the name of image that is used as hyperlinks to detail the contents. Both of them are important components to distinguish the section that summarized the latest information from the other sections that are not essential to be monitored.

Some sections are relatively easy to classify as unimportant sections such as menu column, search subsection, archive column, shopping part and etc. In those sub sections, they contain such common keywords that can be found and we can use them as a value of an attribute for cases in building rules. Some images also can be categorized as a component in a less important part of the Web page documents. Therefore we also collect these images' name to train the rule. In some Web pages, a bold style for main or updated issues is used to enhance the expressiveness that draws user's attention to that section.

Another attribute is a class name which is used in a style sheet. Some designers create arbitrary classes of their own. Style sheets specify text, background colours or some other styles for content in HTML document. Therefore we can check the class name in order to identify the important section in Web pages because they usually define a significant name related to the categories of the section.

Table id is also selected as one of the attributes, since some designers assign particular tables with their own specific ids.

A character or text length is another useful attribute that provides information about the criteria of each section. We also define tables that have no characters except the HTML tags as a table for aligning the structure of Web page layout or the table specified for images only.

The distinctive flavour of web pages is a hyperlink attribute which has the ability to move around from one page to another in the web site or another location. Usually there are fewer hyperlinks in the main content section compared to menu column or other related sections. Thus, we calculate the number of links in each table and put the number as one of the attribute in case within the rules.

We also extract a heading text as one of the attributes. Headings appear larger than the body text. Heading tag is range from <h1> to <h6>. Heading tags are always resided in the main section because it emphasizes the appearance of the text as well as the content.

4.4 Knowledge model

Our training data set has 58 pages with 2,077 tables and testing data set has 7 pages with 199 tables. We did not use all of the collected data because a few of them are not suitable to be a case in training or testing process. We extract the attributes from this data during pre-processing and some of them are detected from a new case during the training process, we then use them to build a knowledge model. We present the specifically defined knowledge model as a case model using RDR tree structure construction. RDR is a special case of decision trees for reasoning with default and it is successfully used to maintain the consistency of the rules created. With RDR, we organize the knowledge base as lists of rules. The system checks the conditions of the main list and sub-list rules until a case is satisfied then the conclusions of the last rule are used.

The knowledge model is built based on a rule type which represents the dependencies between the concepts. We create logical statements that express about an attribute's value of a case and indicate some heuristic relationship between domain expressions. For this reason, we specify a connection between antecedent and the consequent for each rule type. In this study, one case can have more conditions with one conclusion by training a combination of several rules.

A knowledge base contains instances of domain-knowledge types such as concepts, relations and rule types. We only create a manifestation rule for a section monitoring knowledge base. Our knowledge base is not completed during analysis and the knowledge model is still not stable enough. This is shown after we analyse the accuracy of rule's set. Results of the training and testing phase are shown in figure 1.

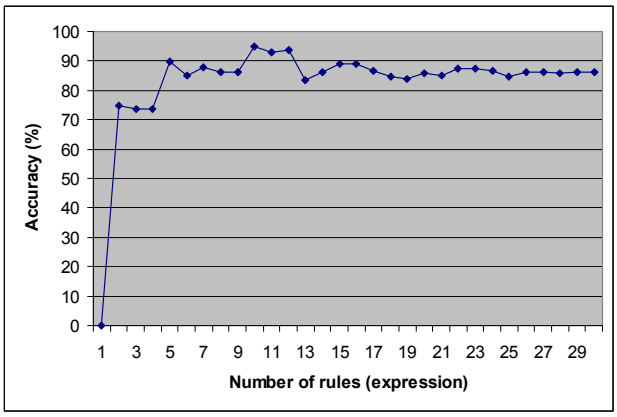

Figure 1. Classification accuracy of section monitoring knowledge base

Given a set of Web documents contain tables; the accuracy of the first rule is about 74.7%. The highest accuracy is 94.8% while 9 rules were tested. The more rules are established, the more stable they are and more cases are satisfied during the final stage of training process. The possibility of the rules to create more incorrect conclusions for testing data is higher if new rules that contain more irrelevant attributes are added. A good case selection for training and testing process is also one of the factors that affect the rule's performance.

5. Conclusion

The result from this study assured that a different way of abstract case can improve classification of Web documents. An abstract case can be presented as a rule type knowledge model that gives an expression about an attribute value of concept. The discussion in this study focuses on Web structure abstraction which is feasible to be used in Web documents classification particularly for section monitoring.

The study shows that the success or the accuracy of a knowledge model may be affected by irrelevant attributes that are used to create an abstract case. It is acknowledged that a good set of cases will lead to a good and compact set of rules. Ripple down rules (RDR) guarantees to offer a simple way of acquiring knowledge and maintains the construction of a rules set. However, the cases used by RDR tend to be historical cases. They sometimes have non-significant attributes that make the knowledge base construction slow and delayed to mature. The conclusion of this study is that the abstract case can improve the classification with the use of a good set of data and significant approach of knowledge acquisition.

In the future implementation, as well as finding a better approach and a good set of cases, we also need to improve the algorithm of data extraction to acquire more relevant attributes in each case. For this scope of study, we found that mathematical and statistical methods are very useful for achieving more features from Web documents.

6. References

[1] H.H. Chen, C.T. Shih, and H.T. Jin, 'Mining Tables from Large Scale HTML Texts', The 18th International Conference, Saarbrucken, Germany, 2000

[2] M.N. Garofalakis, R. Rastogi, S. Seshadri, and K. Shim, 'Data Mining and the Web: Past, Present and Future', Proceedings of WIDM'99, Kansas City, Missouri,1999, pp. 43-47.

[3] M. Hurst, 'Layout and Language: Challenges for Table Understanding on the Web', Proceedings of the First International Workshop on Web Document Analysis (WDA2001), Seattle, Washington, USA. 2002

[4] M. Hurst, 'Classifying TABLE Elements in HTML', Proceeding of the 11th. International World Wide Web Conference (In poster), Honolulu. 2002

[5] J. Palme, J. Why are Most Web Pages Organized in Similar Ways?, 2002, viewed 8 April 2004, <http://dsv.su.se/jpalme/layout/>.

[6] S.S.Park, Y.S. Kim, and B.H. Kang, 'Web Information Management System: Personalization and Generalization', Proceedings on the IADIS International Conference WWW/Internet, Algarve Portugal. 2003

[7] P. Preston, P. Compton, G. Edwards, and B.H. Kang, 'An Implementation of Multiple Classification Ripple Down Rules', Tenth Knowledge Acquisition for Knowledge-Based Systems Workshop, Banff, Canada, SRDG Publications, Department of Computer Science, University of Calgary, Calgary, Canada. 1996

[8] L. Ronsenfeld, and P. Morville, Information Architecture for the World Wide Web, O'reilly, Sebastopol, Canada, 1998

[9] G. Schreiber, H. Akkermans, A. Anjewierden, R.D. Hoog, N. Shadbolt, W.V. Velde, and B. Wielinga, Knowledge Engineering and Management – The Common KADS Methodology, A Bradford Book, Massachusetts, USA, 1999

[10] F. Sebastini, 'Text Classification for Web Filtering', Final POESIA Workshop "Present and Future of Open-Source Content-Based WebFiltering",,Pisa, Italy, 2004

[11] Y. Wang, 'Document analysis: Table Structure Understanding and Zone Content Classification', PhD thesis, University of Washington, 2002

[12] Y.Yang, and S.L. Wo, 'A Framework for Web Table Mining', Workshop On Web Information And Data Management, Virginia USA, 2002

[13] M. Yoshida, K. Torisawa and J. Tsujii, 'A Method to Integrate Tables of the World Wide Web', Proceeding of the 1st International Workshop on Web Document Analysis, Seattle WA, USA, 2001